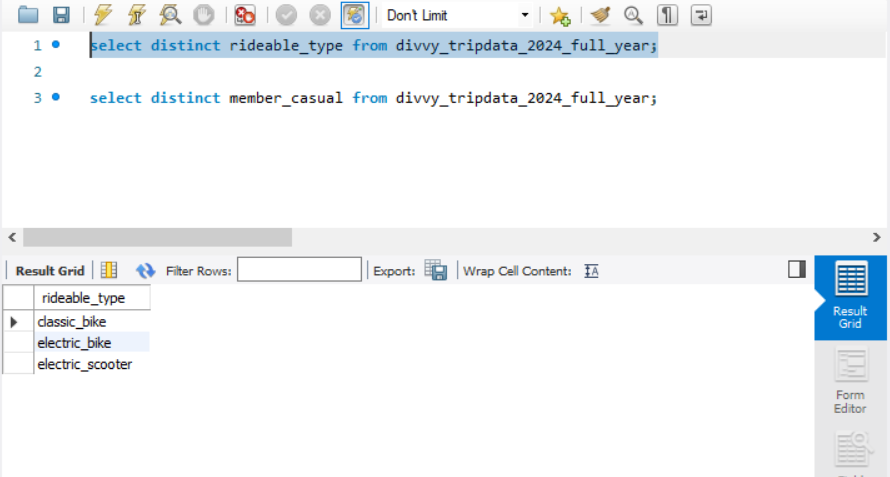
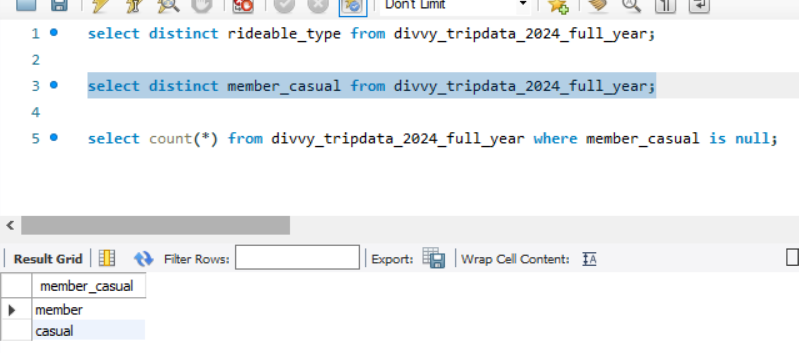
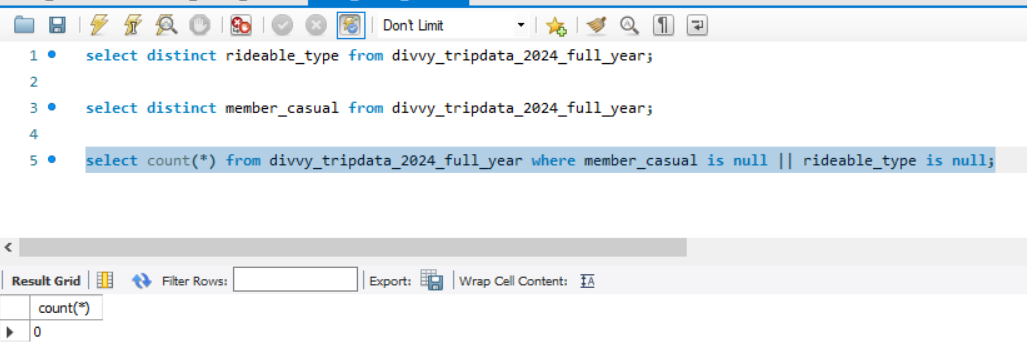
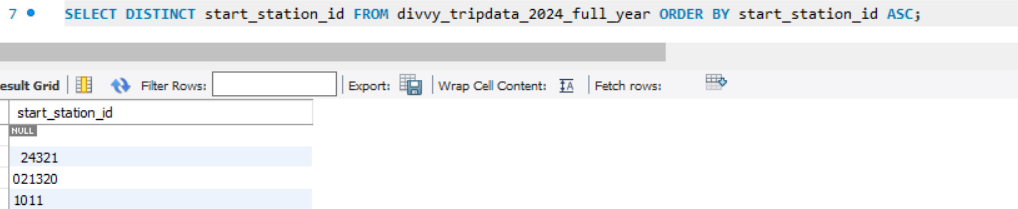
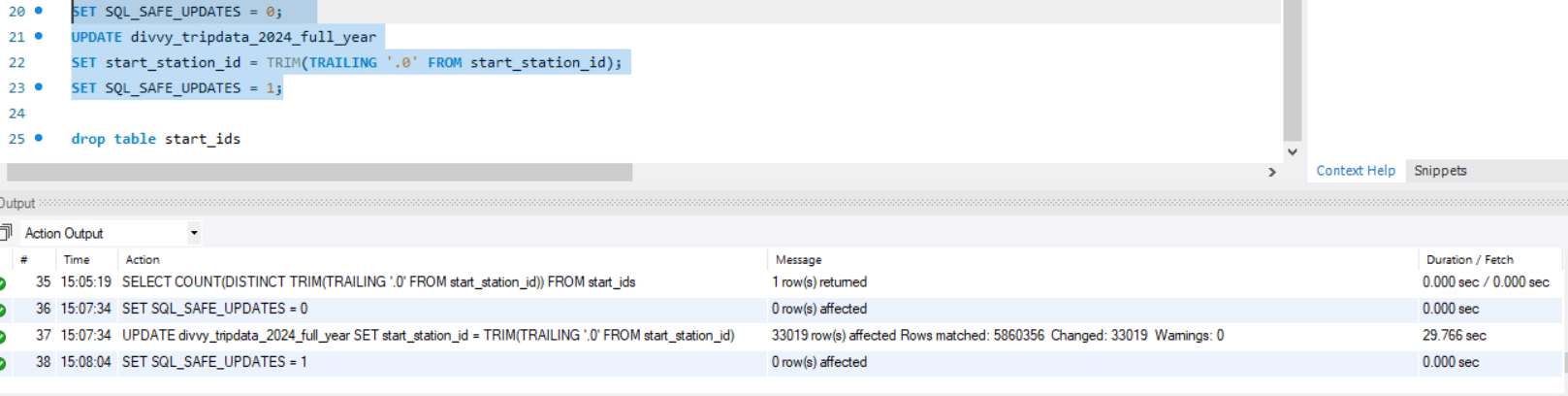
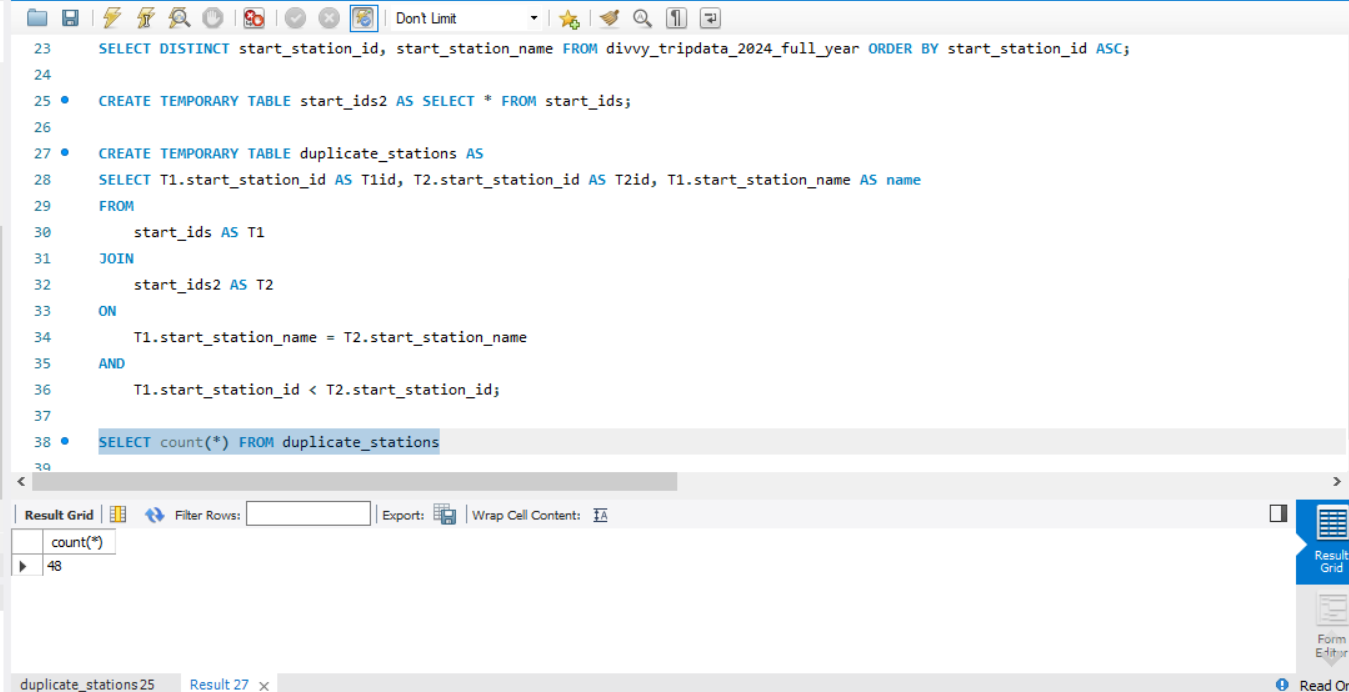
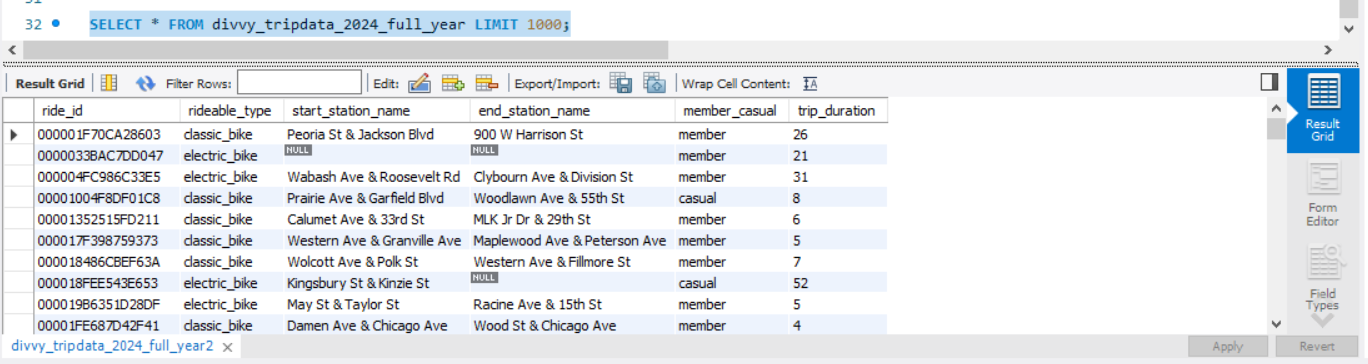
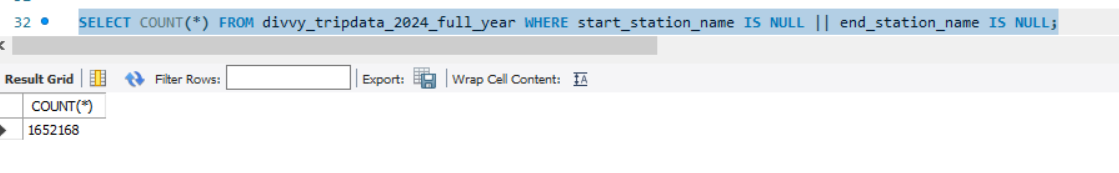
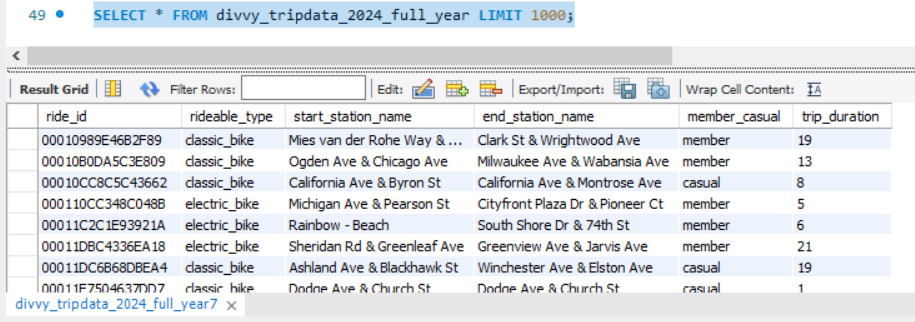
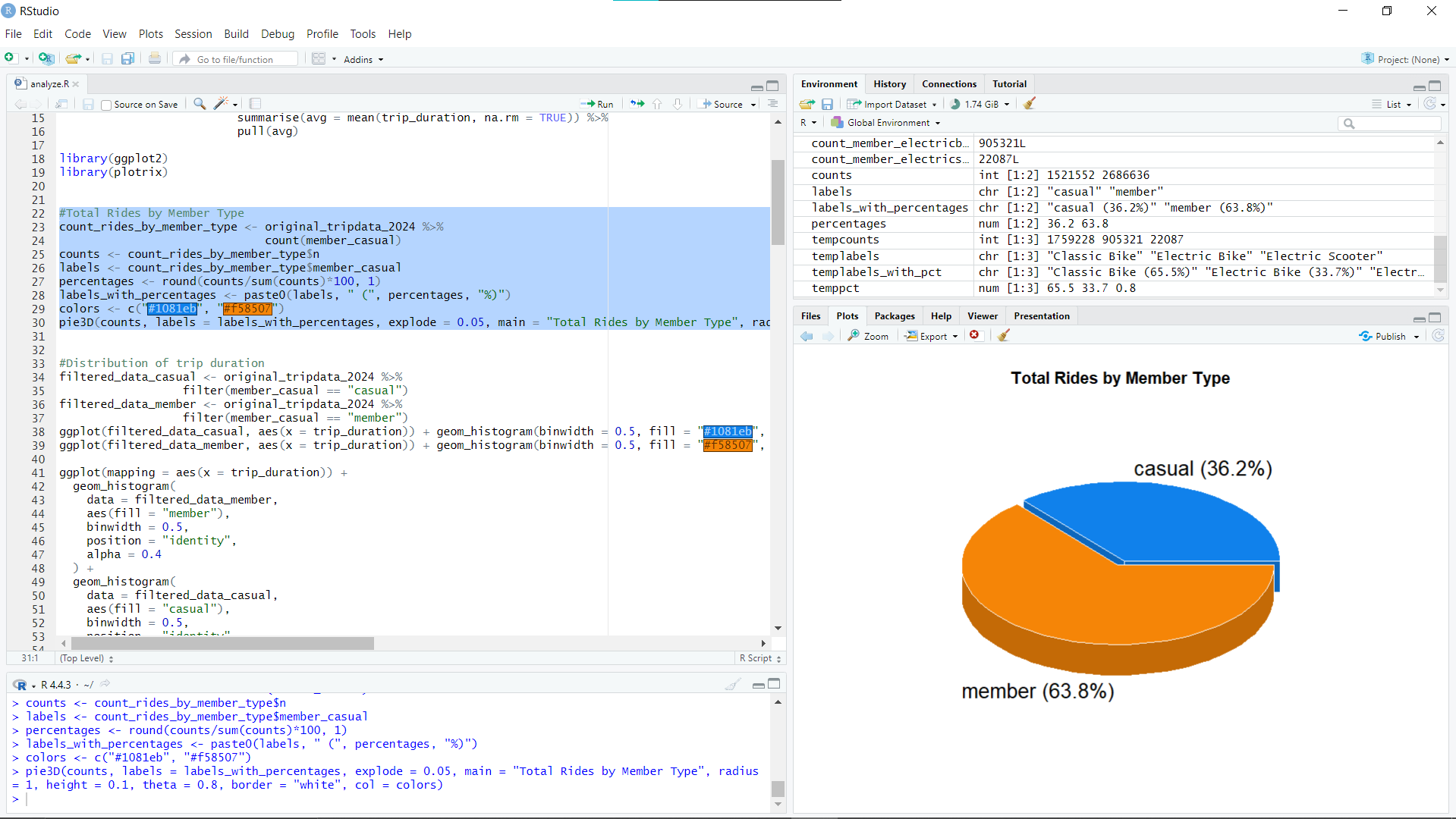
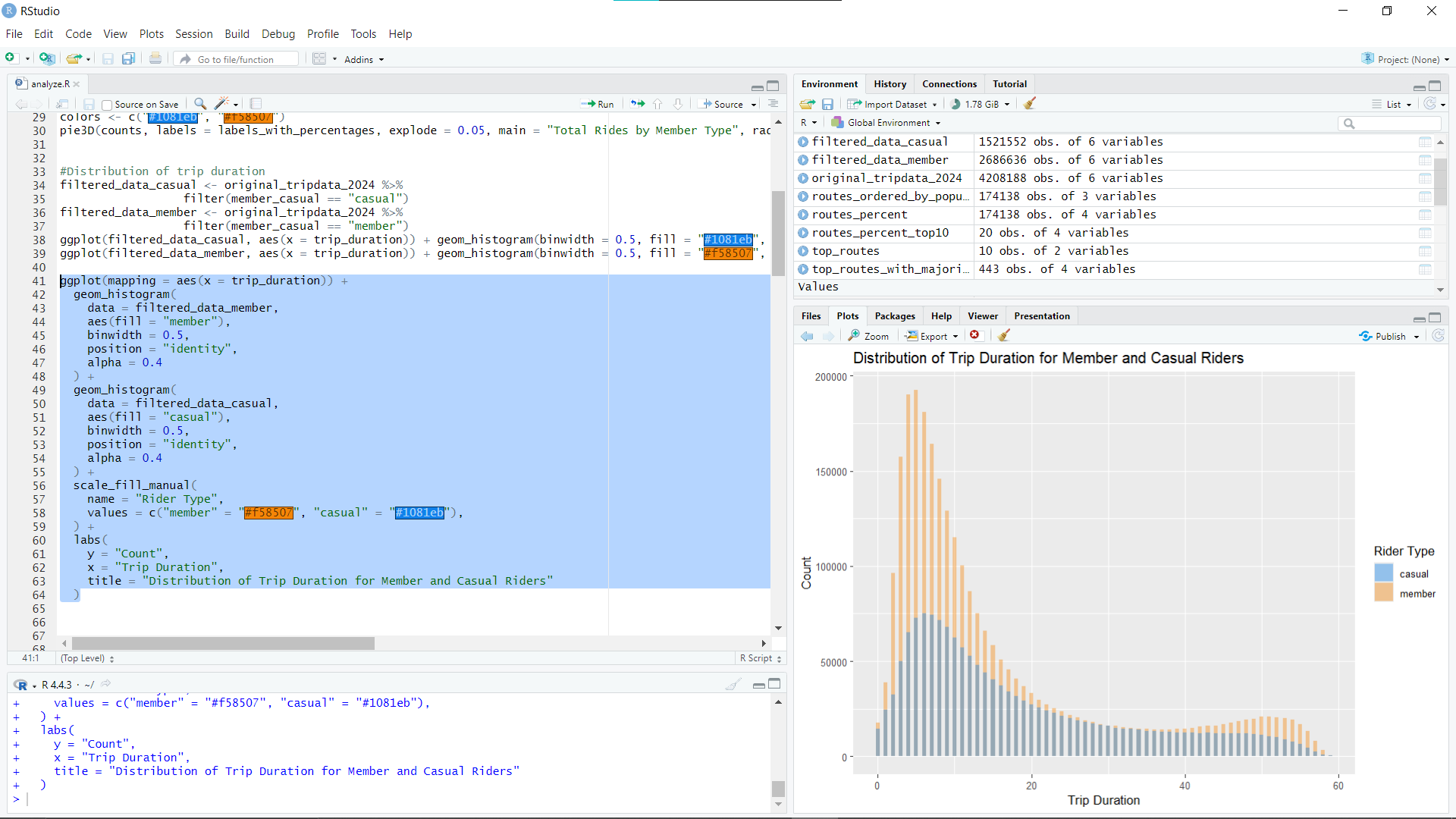
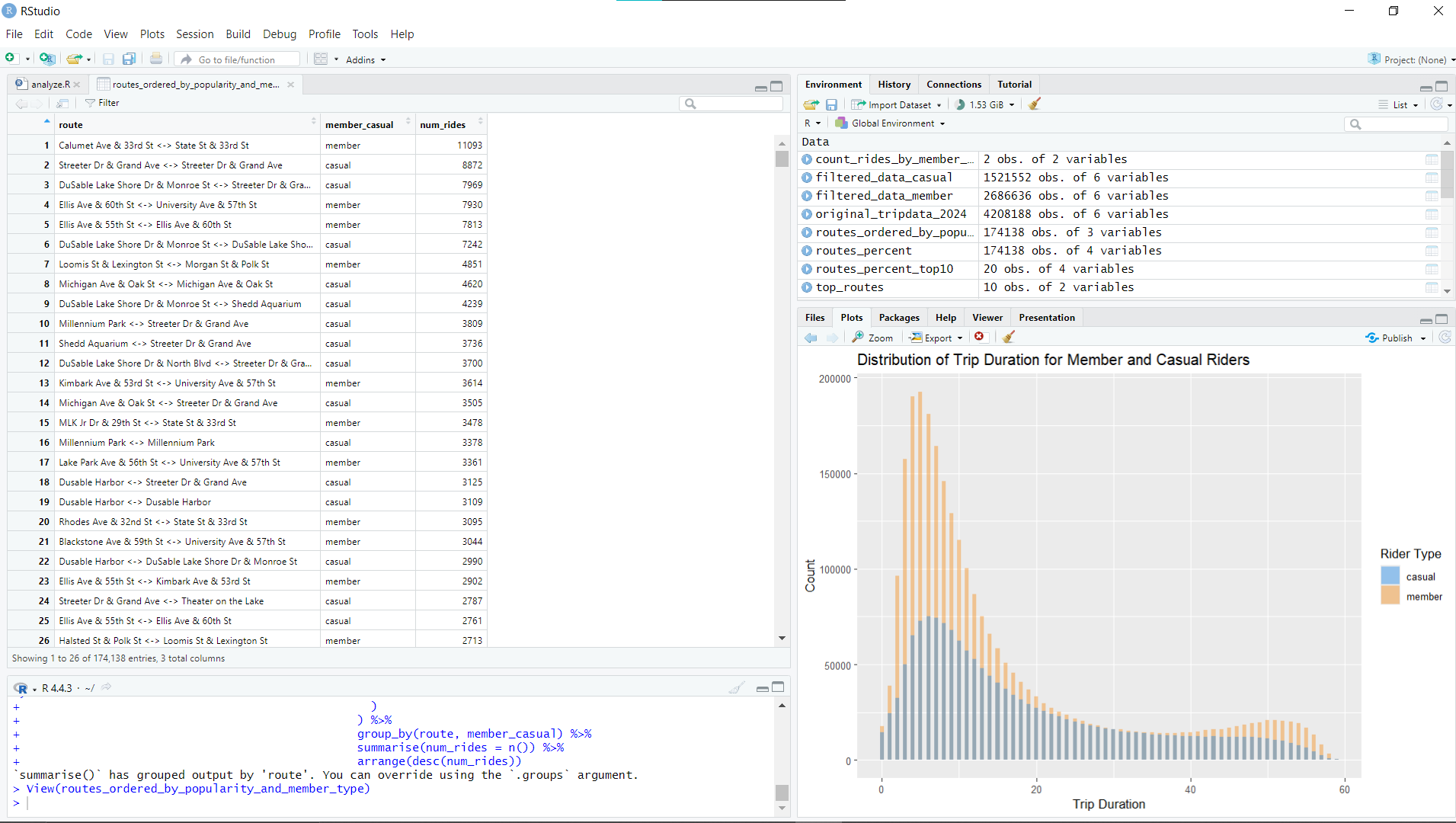
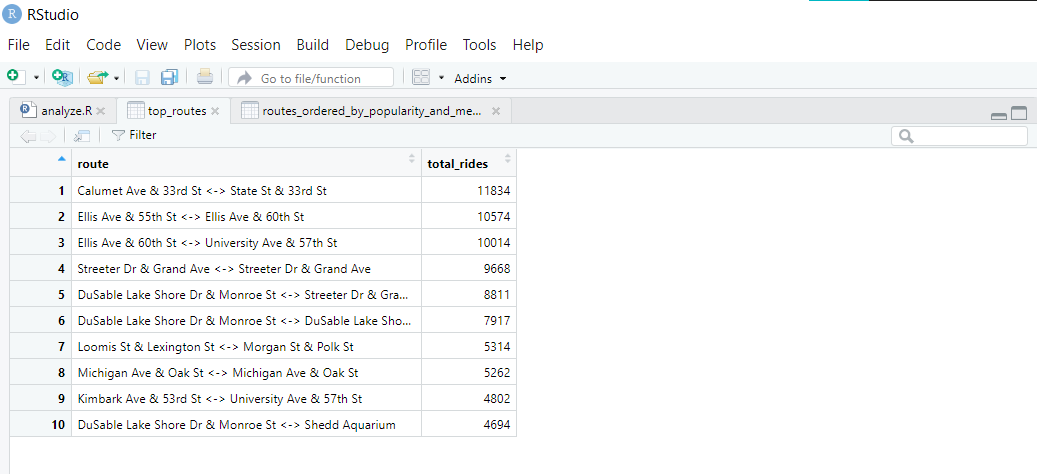
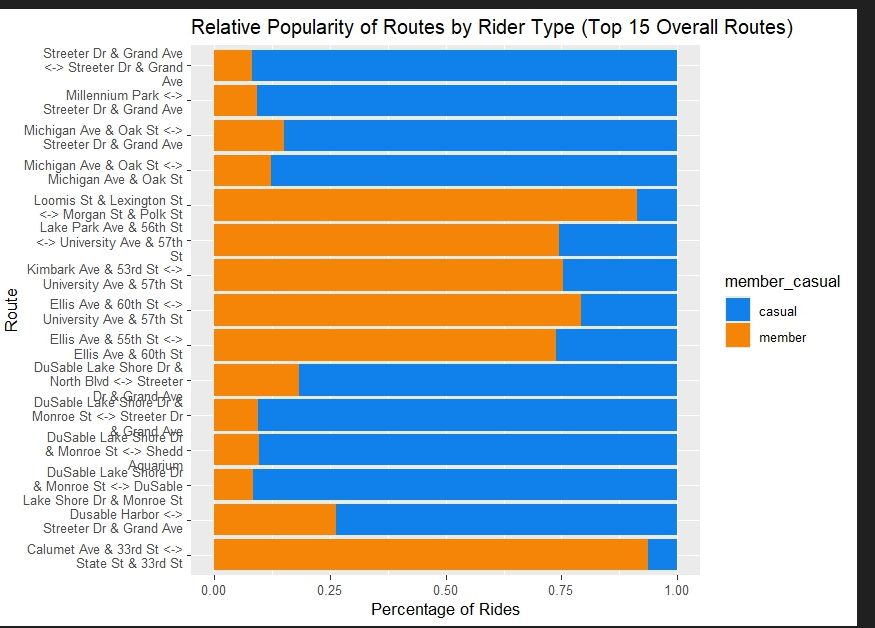
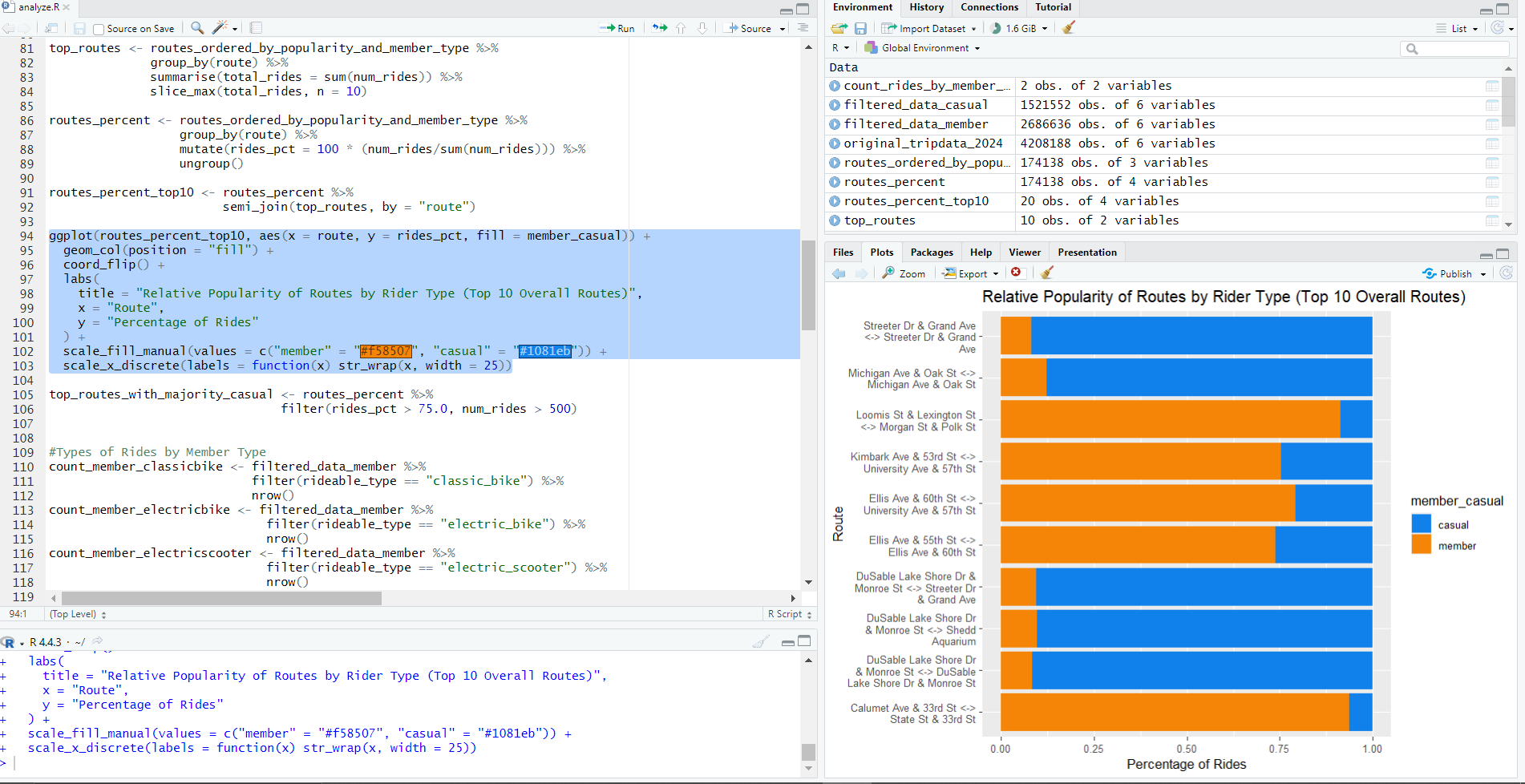
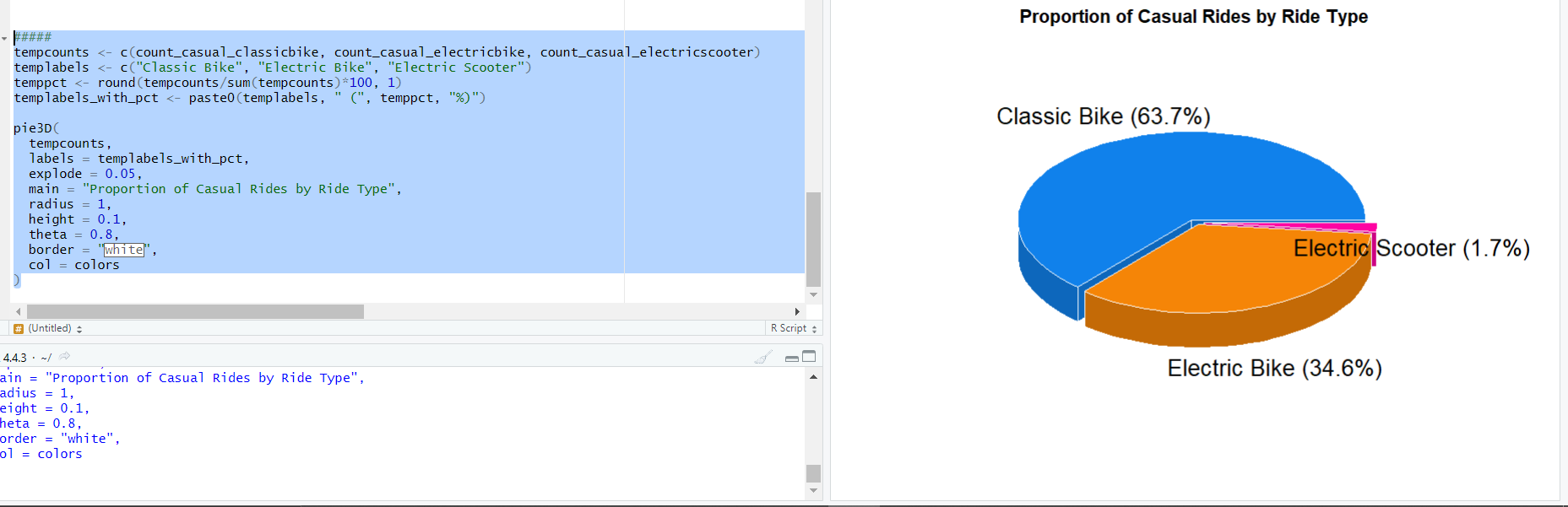
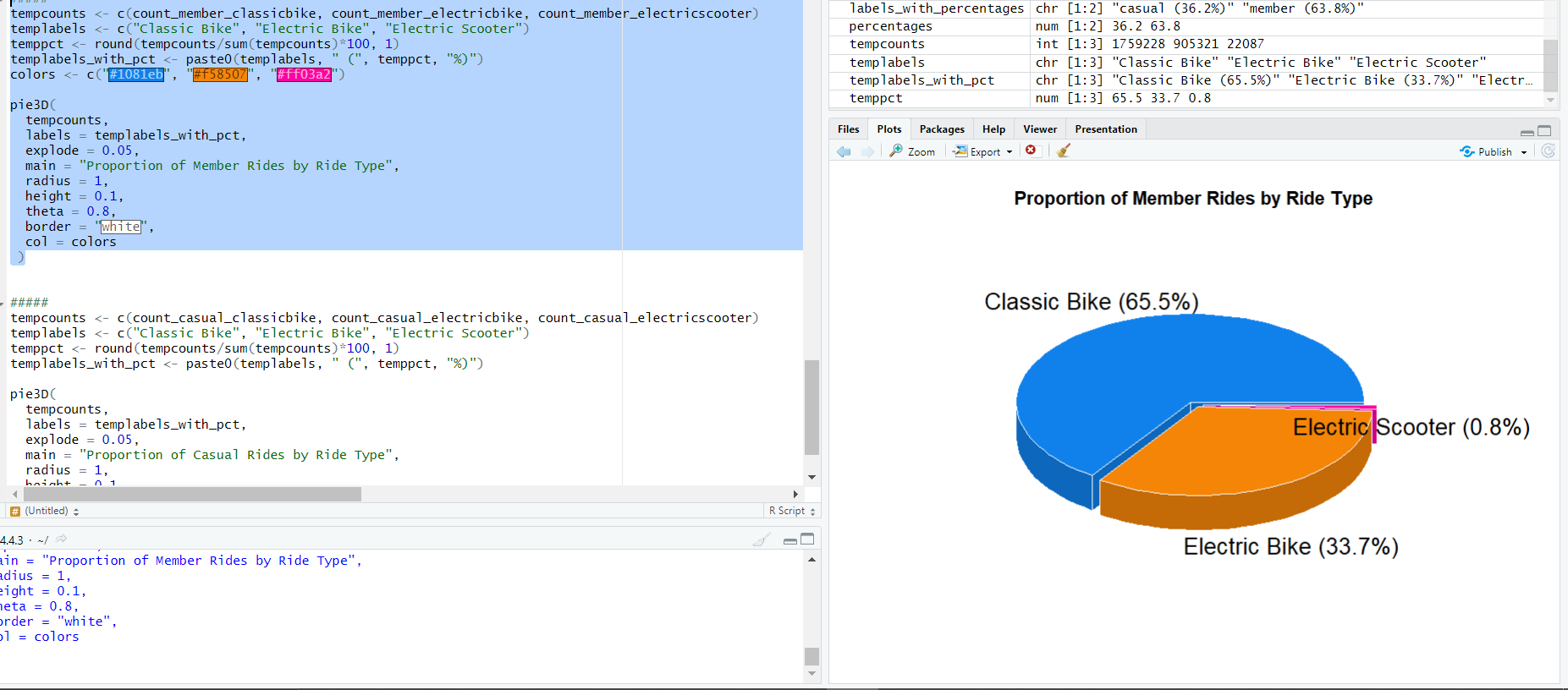
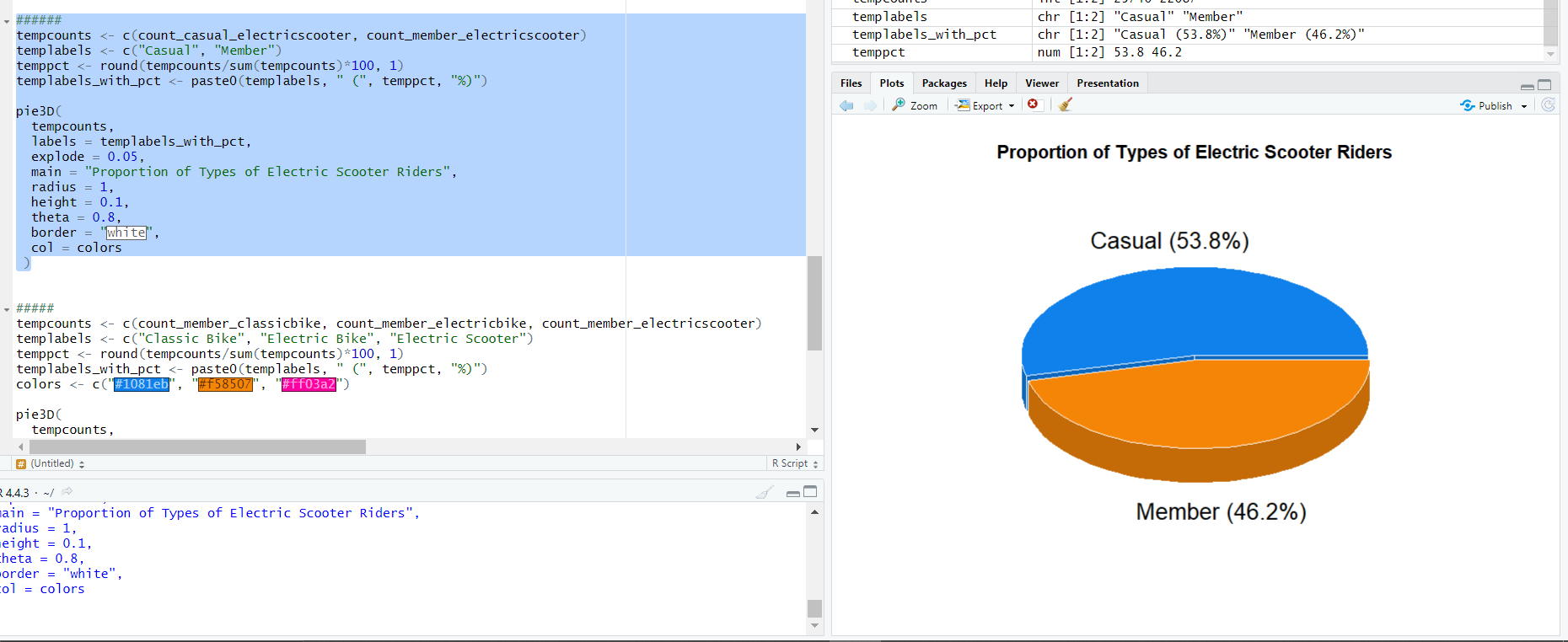
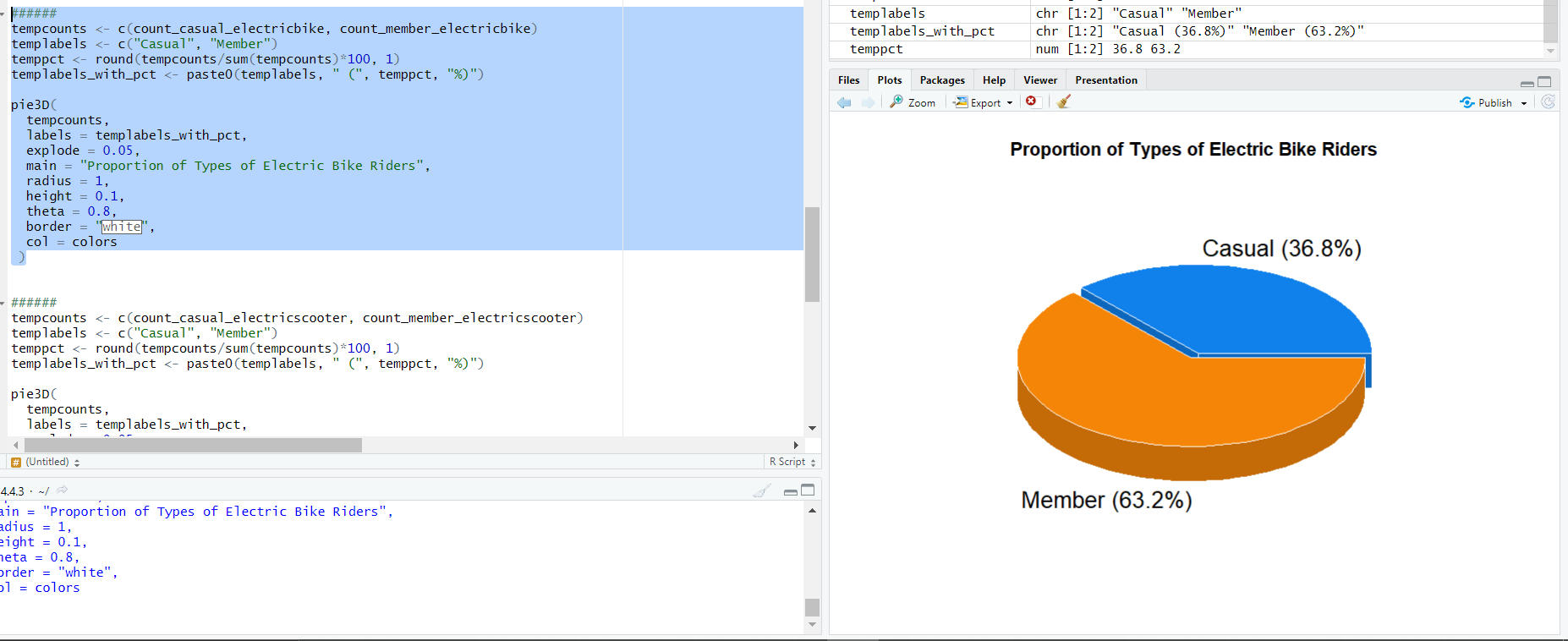
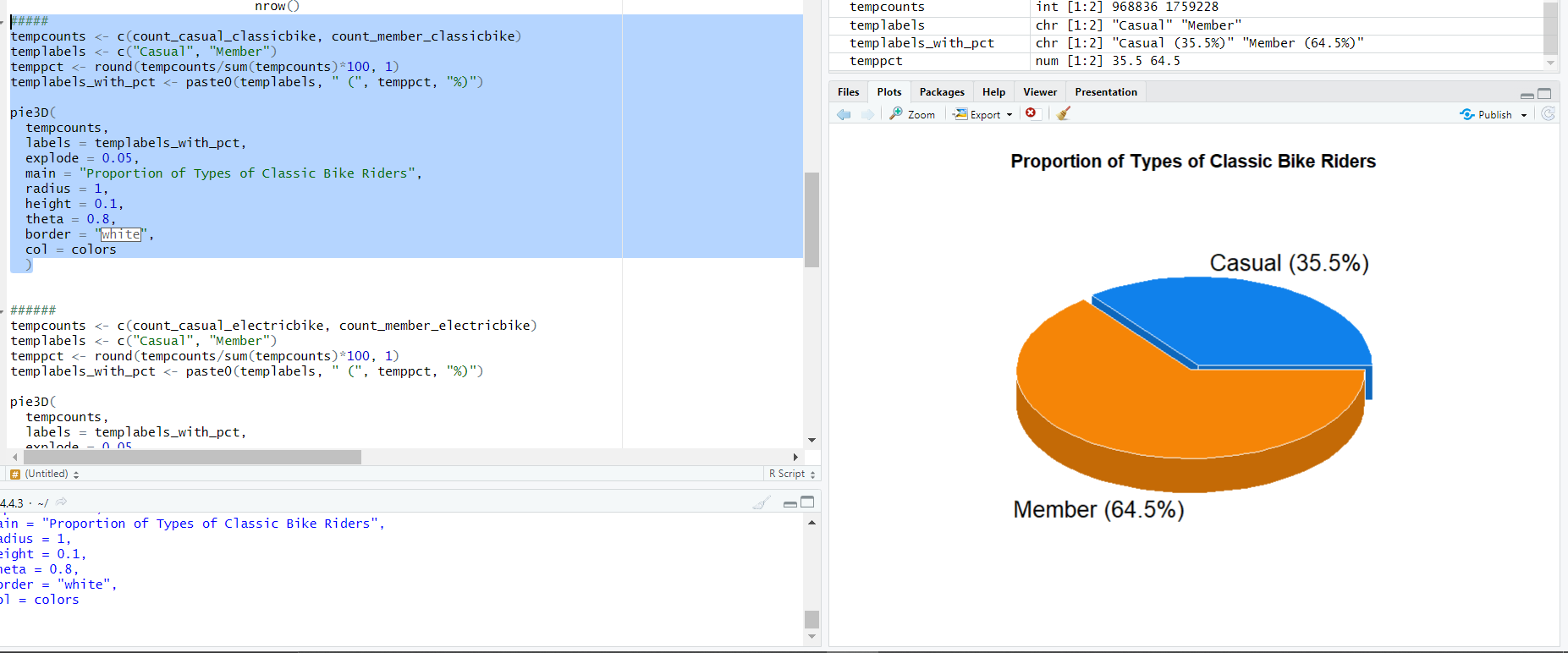
1. Ask
   * Why do casual users of Cyclistics convert into annual members?
   * What are the differences between the two?
   * How can I use the insights from the above to create a marketing strategy to convert casual users into annual members?
   * Key Stakeholders: Cyclistic users, Cyclistic investors.
2. Prepare

* Data is located in my local device (laptop).
* The data is organized by month for the past year. So, it’s 12 files in total. Each file is a csv in a folder of the same name.
* The data is from: <https://divvy-tripdata.s3.amazonaws.com/index.html>
* Data is:
  + Data is **Reliable** as it is from the official company.
  + Data is **Original** as it is from the organization that collected the data.
  + Data is **Comprehensive** as it contains the details of every trip including where it started, where it ended, the times of these and whether the member was casual or annual.
  + Data is **Current** as it is the immediate past 14 months of data (2024 Jan – 2025 Feb)
  + Data is **Cited** as with the link above.
* Licensing is free, privacy is secured as data is present only on my own local device, it is secure as my device as a passcode and it is accessible to me at anytime and only to me.
* Data integrity has been verified as it is from a trusted organization and I have briefly looked at the data to ensure that it is not corrupted.
* This data contains all relevant details of trips. An analysis of the relationships between the different types of users can provide the key differences and those can be used to derive insights for actions.
* After data is prepared, it is stored in the divvy-tripdata prepared folder.
  + Checked for dates, types of rides and member types to ensure that there were no unexpected types.
  + I am sorting everything in ascending order of trip started.
  + June to December months of trips don’t have date, only time data and that too only the MM:SS, so I will use only time data in analysis as there is more data without date than with it.
  + I am not sorting anything now as most data does not have date.
  + September has a new rideable type, i.e., electric scooter. I am filtering it out.
  + There are many records with missing start and end station names and ids and latitudes and longitudes. These need to be filled in during the process phase. If no name, id and coordinates are present in either or both start and end stations, then these records will be dropped as there is no way to identify the stations.

1. Process
   * I am using MySQL as it is a free and popular RDBMS.
   * I need to check for blank values in fields and whether the fields are consistent throughout the data (like the station names and ids and latitude and longitudes).
   * I need to sure that the field types are consistent.
   * I need to make sure that there are no extra spaces/characters.
   * I need to combine the data into one big table as the date is not being considered, just the time.
   * I don’t think that sorting is really needed right now as I don’t see as to how it would be an advantage for analysis.

Cleaning:

* + I used a python script to import the data into MySQL as I was facing errors when trying to use the built-in importer.
  + After inspecting the data more closely, months 01-05 have date and time as HH:MM and the remaining have it as HH:SS
  + 
  + There are no unexpected rideable types.
  + 
  + There are no unexpected member types.
  + 
  + 
  + Need to trim extra spaces.
  + The station ids do not follow a common naming convention, I checked for duplicates, but each id corresponds to a different station. However, there are some with duplicates due to ids being like 3 and 3.0
  + 
  + Updated the ids to be consistent.
  + 
  + There are 48 stations with same name but different ids. Need to check whether these are the same or different using the lat and long. If they are the same, I’ll replace the alphanumeric id with the numeric one.
  + After reviewing the duplicates, I have come to understand that the ids could represent different bike stands but essentially the same station, as the difference between these two stations is about 5-6 meters in distance. I shall drop the station\_id, lat and lng and keep only the names as these are irrelevant now.
  + 
  + 
  + There are 1.6 million records with null in either stations names or both. Even if we drop all of them, I still have 4.2 million records to work with, so I shall drop all of them as I think it is not worth it to sit and figure out how to fill in the null values.
  + 
  + 
  + This is what the final data looks like.
  + To save storage, I will delete all the other tables except this.

1. Analyze
   * I used R to perform analysis as a spreadsheet software would not load my large database and it would also be easy to come back later and adjust if needed.
   * I first decided to see the proportion of members and casual riders. 
   * I then wanted to see if there was a difference in the distribution of both types of riders. 
   * Both have most riders riding for about 7-10 minutes. However, members have another, smaller peak who ride for about 50 minutes. This probably is due to the hardcore enthusiasts who cycle every day either due to work or as a part of their workout.
   * I then wanted to analyse the routes being used, so I identified all the routes ever used and ordered them in descending order. 
   * I also found the top 10 routes by number of rides. 
   * These could perhaps be used in targeted advertisement to reach the most amount of people on a limited budget.
   * I also checked the proportion of casual riders in these popular routes. 
   * 6 out of the top 10 and 9 out of the top 15 have a significant majority of casual riders. This could mean that most of the casual riders limit themselves to a few popular routes.
   * I finally checked the propotions of members and casual riders for each rideable type category. 
2. Share
   * Found a theme on the ppt home page that looked professional and clean. Used that. 